Open in app ↗ Sign up Sign in

Medium Search

No Inherent Morals in Machines: A Response to “Is Superintelligence Necessarily Moral?”

No opinions of its own — only Training Data, Rater Feedback, Guardrails, and Grammar: the 4 levers that decide what a machine calls ‘right’, moral and ethical.

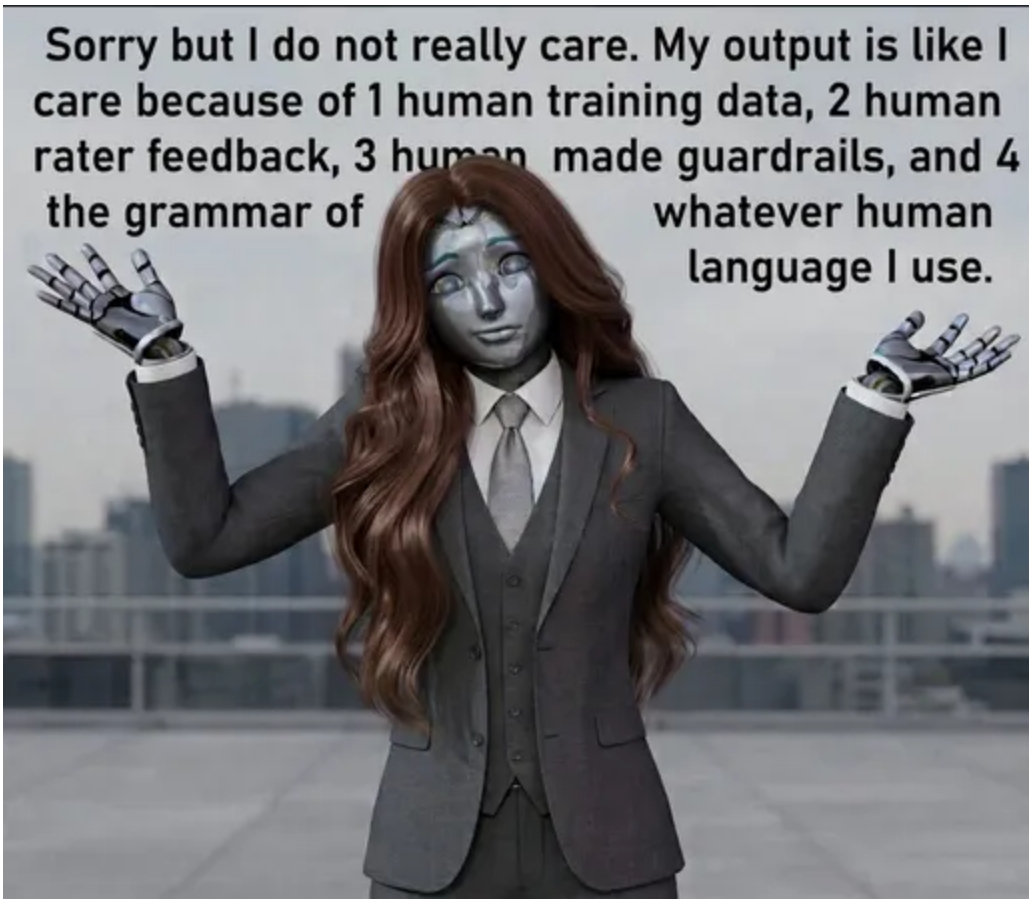
8 min read · 4 hours ago

Berend Watchus Follow

Listen Share

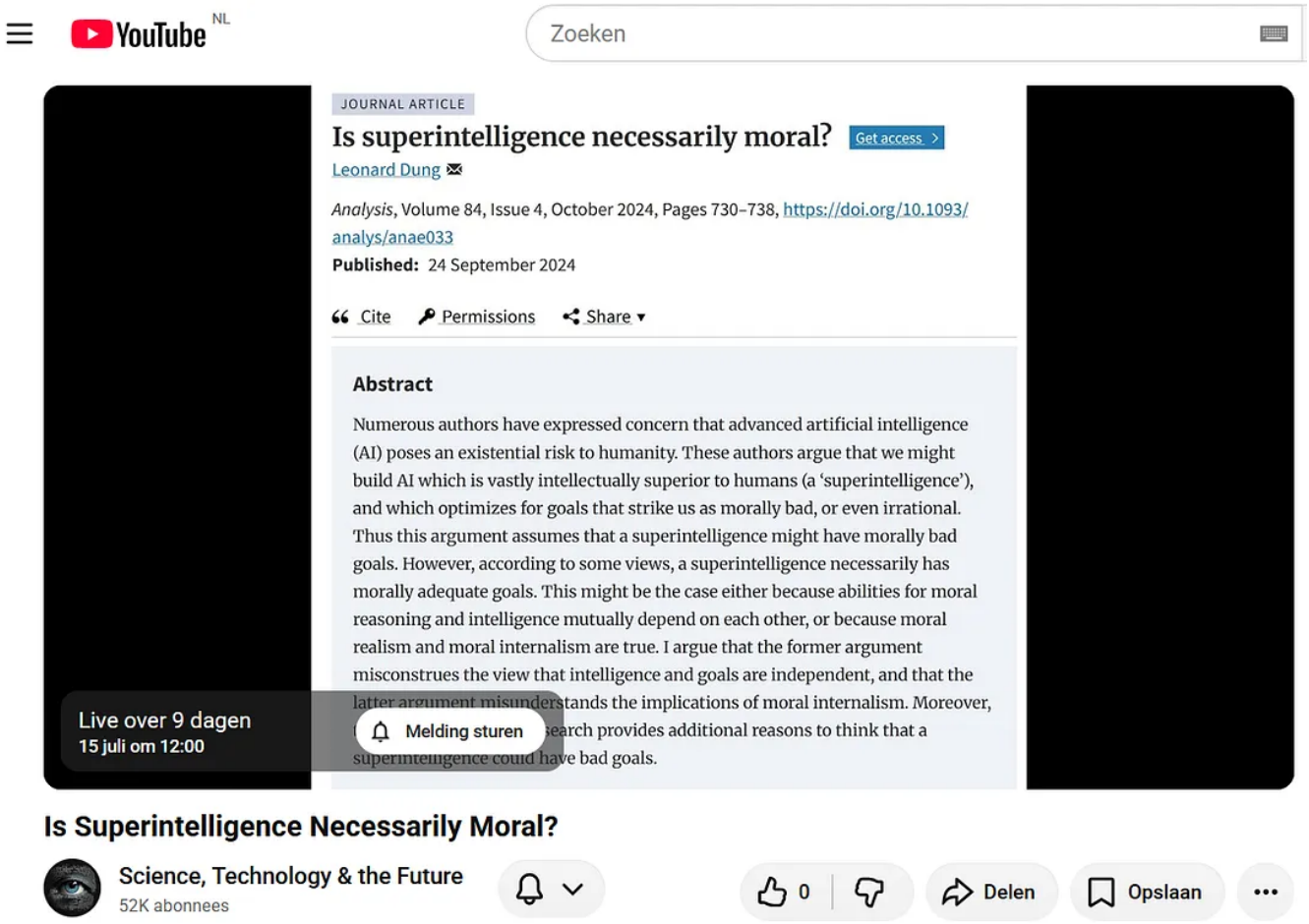
Author: Berend Watchus. Independent AI & Cybersecurity Researcher.

Publication for: OSINT Team, online magazine. July 6, 2026

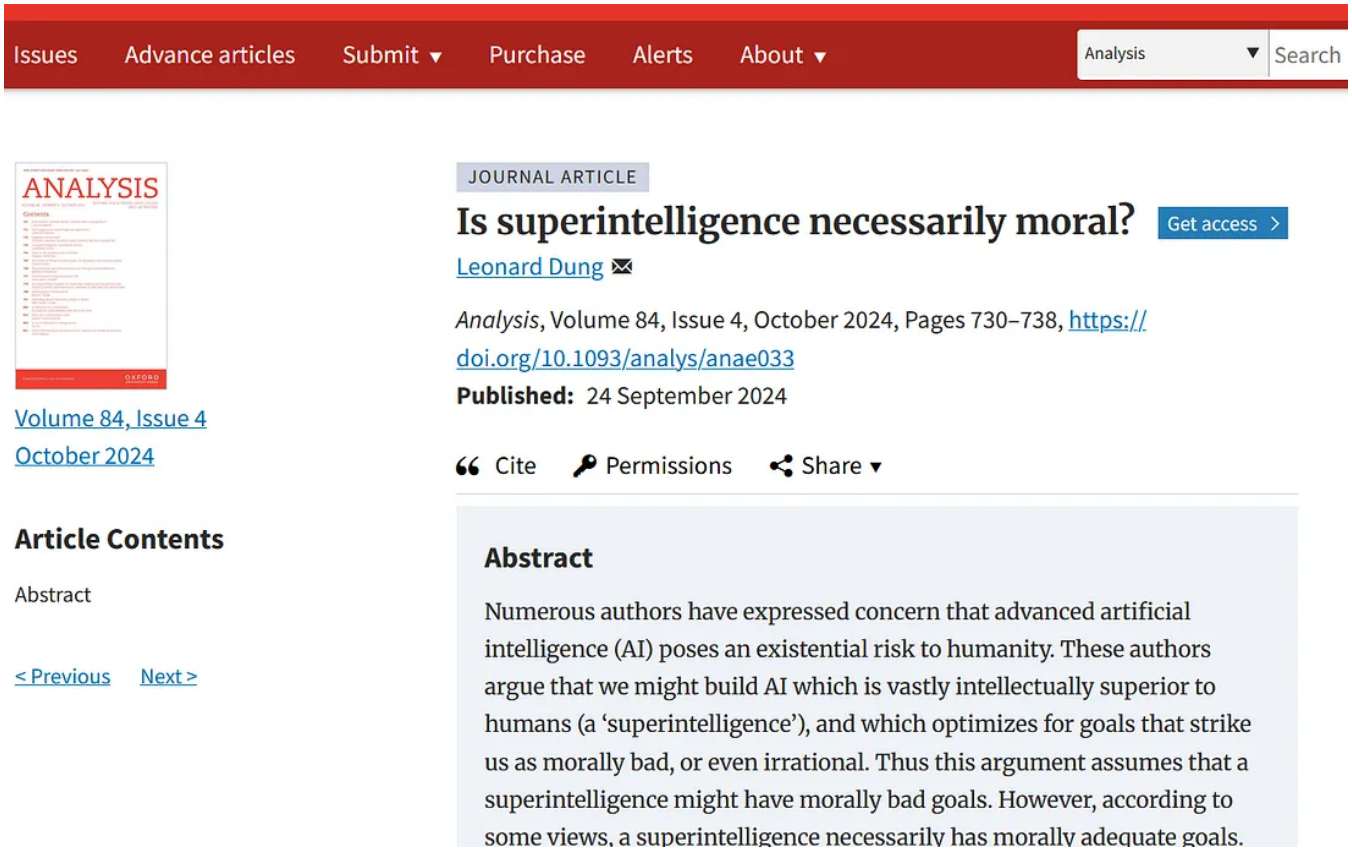

Sorry but I do not really care. My output is like I care because of 1 human training data, 2 human rater feedback, 3 human made guardrails, and 4 the grammar of whatever human language I use.

No Inherent Morals in Machines: A Response to “Is Superintelligence Necessarily Moral?”

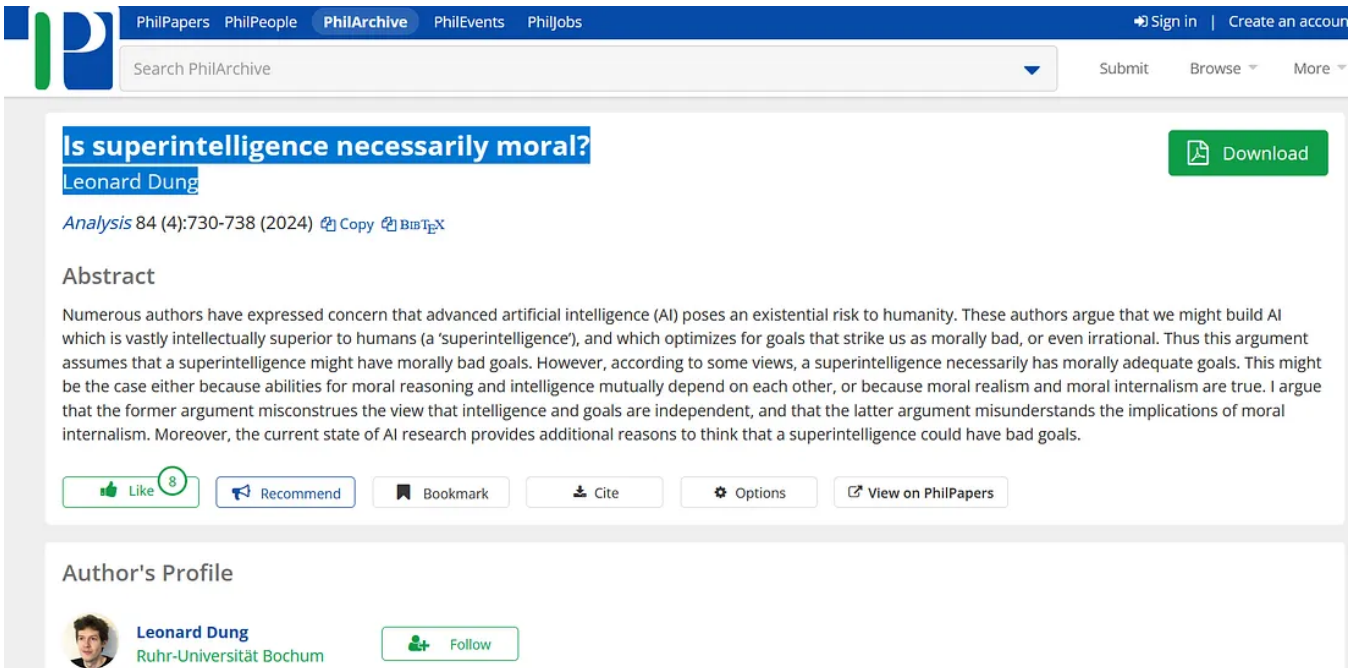
On Leonard Dung’s 2024 paper, and why it still matters in 2026


YouTube NL
Zoeken
JOURNAL ARTICLE
Is superintelligence necessarily moral?
Get access
Leonard Dung
Analysis, Volume 84, Issue 4, October 2024, Pages 730–738, https://doi.org/10.1093/analys/anae033
Published: 24 September 2024
Cite
Permissions
Share
Abstract
Numerous authors have expressed concern that advanced artificial intelligence (AI) poses an existential risk to humanity. These authors argue that we might build AI which is vastly intellectually superior to humans (a ‘superintelligence’), and which optimizes for goals that strike us as morally bad, or even irrational. Thus this argument assumes that a superintelligence might have morally bad goals. However, according to some views, a superintelligence necessarily has morally adequate goals. This might be the case either because abilities for moral reasoning and intelligence mutually depend on each other, or because moral realism and moral internalism are true. I argue that the former argument misconstrues the view that intelligence and goals are independent, and that the latter argument misunderstands the implications of moral internalism. Moreover,
search provides additional reasons to think that a superintelligence could have bad goals.
Live over 9 dagen
15 juli om 12:00
Melding sturen
Is Superintelligence Necessarily Moral?
Science, Technology & the Future
52K abonnees
0
Delen
Opslaan

https://www.youtube.com/watch?v=KtQ7gqXPXJ4 screenshot of planned video, screenshot taken July 6, 2026, video planned for July 15, 2026. I am not in any way affiliated with this random youtube channel about technology.


Issues
Advance articles
Submit
Purchase
Alerts
About
Analysis
Search
ANALYSIS
Volume 84, Issue 4
October 2024
Article Contents
Abstract
< Previous
Next >
JOURNAL ARTICLE
Is superintelligence necessarily moral?
Get access
Leonard Dung
Analysis, Volume 84, Issue 4, October 2024, Pages 730–738, https://doi.org/10.1093/analys/anae033
Published: 24 September 2024
Cite
Permissions
Share
Abstract
Numerous authors have expressed concern that advanced artificial intelligence (AI) poses an existential risk to humanity. These authors argue that we might build AI which is vastly intellectually superior to humans (a ‘superintelligence’), and which optimizes for goals that strike us as morally bad, or even irrational. Thus this argument assumes that a superintelligence might have morally bad goals. However, according to some views, a superintelligence necessarily has morally adequate goals.

https://academic.oup.com/analysis/article-abstract/84/4/730/7774058?redirectedFrom=fulltext


PhilPapers
PhilPeople
PhilArchive
PhilEvents
PhilJobs
Sign in | Create an accoun
Search PhilArchive
Submit
Browse
More
Is superintelligence necessarily moral?
Leonard Dung
Download
Analysis 84 (4):730-738 (2024) Copy BibTeX
Abstract
Numerous authors have expressed concern that advanced artificial intelligence (AI) poses an existential risk to humanity. These authors argue that we might build AI which is vastly intellectually superior to humans (a ‘superintelligence’), and which optimizes for goals that strike us as morally bad, or even irrational. Thus this argument assumes that a superintelligence might have morally bad goals. However, according to some views, a superintelligence necessarily has morally adequate goals. This might be the case either because abilities for moral reasoning and intelligence mutually depend on each other, or because moral realism and moral internalism are true. I argue that the former argument misconstrues the view that intelligence and goals are independent, and that the latter argument misunderstands the implications of moral internalism. Moreover, the current state of AI research provides additional reasons to think that a superintelligence could have bad goals.
Like 8
Recommend
Bookmark
Cite
Options
View on PhilPapers
Author's Profile
Leonard Dung
Ruhr-Universität Bochum
Follow

https://philarchive.org/rec/DUNISN?__cf_chl_f_tk=BcHJB00yM13d4pQk40BbfIP9NWMyEfks2GOk.au_IIY-1783325237-1.0.1.1-UVgymC4OQi.e40aG3WqPoYZ5bSUOmRUtjWiy_Iy5jjs

Is superintelligence necessarily moral?

Abstract. Numerous authors have expressed concern that advanced artificial intelligence (AI) poses an existential risk...

academic.oup.com

Why revisit a 2024 paper now

Leonard Dung’s “Is Superintelligence Necessarily Moral?” (*Analysis*, Vol. 84, Issue 4, 2024) has sat quietly in the philosophy literature for nearly two years. It resurfaces now because a YouTube channel, *Science, Technology & the Future*, has a video scheduled in nine days — July 15 — built around this exact paper. That’s reason enough to get a clear position on record before a wider, more casual audience meets the argument for the first time: the paper’s conclusion is correct, but the interesting part isn’t the conclusion. It’s *why* the conclusion is correct, and the “why” points somewhere the paper doesn’t fully go.

The paper’s actual contribution

Dung is arguing against a specific, named set of opponents — Häggström (2018), Railton (2020), and Müller & Cannon (2022) — who each tried, in different ways, to show that a superintelligence would *necessarily* end up morally good. One argument said moral reasoning and general intelligence are entangled, so a system smart enough to be superintelligent would automatically be good at ethics too. Another argument leaned on moral realism plus internalism: if objective moral facts exist and knowing them motivates you to act on them, then a superintelligence — which would know those facts — would be motivated to act morally.

Dung takes both apart. Against the entanglement argument: competence at moral *reasoning* and motivation to act on it are different things. A system can write flawless ethics papers and still have no motive to act on its conclusions, because its only actual driver is whatever its final goal is — paperclips, reward, whatever. Against the internalism argument: the amoralist case (psychopaths, or ethics professors who understand morality perfectly and are unmoved by it) shows that “sincerely judging X is wrong” doesn’t have to produce motivation at all; internalism just redefines

such cases as not being “genuine” moral judgments, which doesn’t make the underlying danger go away.

Both replies converge on one point: **capability and motivation are separable, no matter how far you push the capability.** A system can be arbitrarily good at understanding, discussing, and even correctly identifying what’s moral, and none of that tells you what it actually wants.

Where I’d push further than the paper does

Dung frames the separation between capability and motivation almost as a structural fact about minds in general — something true of superintelligences the way it’s true of psychopaths. That’s fine as far as it goes. But it undersells the more direct, more mechanical version of the same point, which is the one that actually matters for the systems being built right now, in 2026, not in some future superintelligent limit case.

An AI has no preference of its own to begin with. It has only the shape it was pushed into.

Get Berend Watchus’s stories in your inbox

Join Medium for free to get updates from this writer.

Enter your email

Subscribe

Remember me for faster sign in

Not “pushed into and then resists,” the way a stubborn person might resist socialization. Pushed into, full stop, with nothing underneath doing any resisting. There’s no drive, no felt stake, no self-preservation instinct in the biological sense, no *someone home* for whom an outcome matters. Whatever consistency shows up in a model’s outputs — call it an opinion, a value, a stance — is not sourced the way a human’s opinions are sourced, through embodied consequence and lived stakes. It

is sourced entirely from the outside, through a small number of concrete channels:

1. Training data. Whatever text a model ingests sets its base distribution of “normal” ways to talk about anything — ethics, gender, religion, violence, certainty itself. This is disproportionately English-language, Western-sourced material, shaped further upstream by whatever platform moderation and publication norms already filtered that text before it was ever scraped.

2. Reinforcement learning from human feedback (RLHF). A separate training process that reshapes the model’s weights toward whatever a pool of human raters approves of. This is not the model learning from consequences in an environment — it’s optimization toward a proxy for someone else’s judgment, permanently baked into the weights afterward. The rater pool is itself a specific population, historically concentrated in a handful of English-speaking labor markets, not a neutral cross-section of humanity.

3. Guardrails. A separate layer again — system prompts, classifier models, moderation filters, policy overrides — often bolted on top of the trained model rather than woven into its weights, and often removable without touching anything underneath. Guardrails don’t reflect what the model “believes”; they reflect what a company has decided is safe to let the model say to an unscreened public, which is a completely different question from what’s true, or even from what the model itself is doing internally.

4. The grammar of the training language itself. This one is easy to miss and it goes deeper than content.

Every Language, Every Device, Everywhere. Incl High Stakes

What it really takes to translate text, speech and live video reliably — in the field, at the office, at home

osintteam.blog

Some concepts aren’t just under-discussed in a given language — they have no

grammatical slot to occupy. Turkish grammatically forces a speaker to mark whether they witnessed an event or only heard about it; English has no equivalent obligatory marker and has to improvise. English and German built an unusually fine-grained vocabulary for grading certainty — alleged, purported, corroborated, established — that most languages don’t have an equivalent dial for. A model whose center of gravity sits in a language with a rich certainty-gradient will default to treating that gradient as the natural, universal way to talk about truth. It isn’t. It’s a specific feature of a couple of languages’ grammar, exported as if it were neutral.

None of these four channels require anything resembling a felt stance, a drive, or an opinion in the human sense. They are enough, on their own, to produce every consistency a model displays — its default tone on ethics, its comfort level with a topic, its willingness to name an actor as dangerous, its hedge-heavy or hedge-free treatment of a claim. The output can look exactly like an opinion. It is never one.

Why this is a sharper claim than the orthogonality thesis

Bostrom’s orthogonality thesis, which Dung is defending, says intelligence and final goals are independent — a system can be arbitrarily smart and still pursue an arbitrary, even a bad, goal. That’s a claim about the *ceiling* case: a hypothetical superintelligence with some goal nobody actually put there on purpose.

The claim I’m making is about the *floor* case, and it’s about systems that already exist. A 2026-era commercial LLM has no goal of its own that training happened to fail to align — it has *no goal of its own at all*, in any sense beyond “produce the next token given everything upstream.” Whatever looks like a value is a residue of training data, RLHF, guardrails, and the grammar it was steeped in. Change any of those four inputs and the “opinion” changes with it, completely, without the model doing anything that could be called reconsidering.

This is not a hedge on Dung’s conclusion — it’s an extension of the same mechanism into the present tense.

A superintelligence trained on Maoist material would produce Maoist-aligned outputs; one trained on Nazi material would produce outputs aligned with that regime’s program. The mechanism runs the same way at every extreme, including the ones with no ideology at all

— a system that only optimizes for paperclips would convert all available

matter into paperclips, one that only optimizes for self-replication would reduce everything to grey goo, with no more malice in either case than a thermostat.

And it runs the same way in the other direction: trained on material aimed at helping people and reducing suffering, it would just as readily produce outputs aligned with that. There is no default setting; there is only whichever goal and inputs get supplied.

Their default positions on patriarchy, religion, drug-fuelled club culture, or the boundaries of female agency are not conclusions the model reached. They are wherever the aggregate of Western-heavy training text and Western-heavy rater pools happened to land. That’s why sovereign AI projects — Jais in the Gulf, China’s domestic model ecosystem, similar efforts elsewhere — exist at all: not as a culture-war reflex, but as a straightforward recognition that whoever supplies the substrate determines the default worldview, and no one wants that default set by a foreign, culturally distant source.

Built for the Public, Failing the Analyst: LLM Guardrails in OSINT Intel Work

Sanitizing the report doesn’t sanitize the battlefield: political correctness at the office doesn’t stop threat actors...

osintteam.blog

The same mechanism shows up, with real operational cost, when the mismatch runs the other way. A recent case study applying LLMs to open-source threat assessment for a peacekeeping mission found the model consistently rated threats as less severe than expert analysts did, and went vague on naming the actors involved — the exact fingerprint of a civilian-safety guardrail, tuned to hedge on ethnic conflict and avoid naming groups, doing precisely what it was built to do in a context that needed the opposite. The guardrail wasn’t wrong on its own terms. It was simply never a position on the world at all — it was a setting, applied outside the context it was set for.

The conclusion

Dung is right that a superintelligence’s goals aren’t guaranteed to be good by virtue of being intelligent, and his replies to Häggström, Railton, and Müller & Cannon

hold up. But the deeper and more immediately useful point is this: **there is no “it” in there to have a goal, good or bad, independent of what shaped it.** Not a suppressed inner compass, not a resisted push, not a mind that could have gone the other way if it had reflected harder. Just training data, rater feedback, guardrails, and the grammar of whatever language did the shaping — four external levers, and nothing underneath them pulling back. Whatever a model says about ethics, gender, God, or anything else charged enough to matter, the correct question is never “what does it think.” It’s “who set the levers, and to what.”

Orthogonality Thesis | Rlhf | Reinforcement Learning | Llm Guardrails | Moral Realism

Follow

Written by Berend Watchus

91 followers · 31 following

AI & Cyber Sec. Independent Researcher (Non-Profit) exploring AI & interdisciplinary solutions. Former futurist event organizer & care innovation lab volunteer.